ISQS 6349: PREDICTIVE ANALYTICS

Dr. Paige S. Rutner

Midterm Project

Predict Blood Donations

Created by:

Rushan Shakya

Tahrima Mustafa

Cam Moy

Nov 21, 2017

Table of Contents

|  |  |
| --- | --- |
| **Introduction** ………………………………………………………………………….. | 3 |
| Blood Donation and Data Science ………………………………………………. | 3 |
|  |  |
| **Literature** ……………………………………………………………………………… | 3 |
| Big data analytics in healthcare: promise and potential | 4 |
|  |  |
| Prediction of hemoglobin levels in whole blood donors: how to model donation history ……………………………………………………………………….. | 5 |
|  |  |
| Predicting blood donation intentions and behavior among Australian blood donors: testing an extended theory of planned behavior model ………………… | 6 |
| **Data Mining/Cleaning** ………………………………………………………………. | 7 |
|  |  |
| **Data Visualization** …………………………………………………………………… | 9 |
| Univariate Visualization of the Variables ………………………………………... | 10 |
|  |  |
| Multivariate Visualization of the Variables ……………………………………… | 15 |
|  |  |
| **Predictive Analysis** …………………………………………………………………. | 19 |
| Logistic Regression ……………………………………………………………….. | 19 |
|  |  |
| Support Vector Machine ………………………………………………………….. | 24 |
|  |  |
| **Discussion and Recommendations** ……………………………………………… | 27 |
| Analysis for Logistic Regression and SVM Models ……………………………. | 27 |
|  |  |
| Comparing the Logistic Regression and SVM Models ………………………… | 31 |
|  |  |
| Recommendation ………………………………………………………………….. | 32 |
|  |  |
| **Driven Data Submission** …………………………………………………………… | 33 |
|  |  |
| **References** …………………………………………………………………………… | 34 |

:

Table of Figures

Figure 1 Boxplots for possible predictors 7

Figure 2 Graph showing the relationship within the original data 9

Figure 3 Graph showing the relationship with the new added variables 10

Figure 4 Histogram for Months since last donation 11

Figure 5 Histogram for number of donations 12

Figure 6 Histogram for total volume donated 12

Figure 7 Histogram for Months since first donation 13

Figure 8 Histogram for Made donation in March 2007 13

Figure 9 Multivariate histogram for Months since first donation 15

Figure 10 Multivariate histogram for months since last donation 16

Figure 11 Multivariate histogram for number of donations 17

Figure 12 Multivariate histogram for Average donation per period 17

Figure 13 Multivariate histogram for Ml-MF ratio 18

Figure 14 Multivariate histogram for Made Donation in March 2007 19

Figure 15 Residual Analysis for Logistic Regression (Residuals vs Fitted Values) 27

Figure 16 Residual Analysis for Logistic Regression (Standardized Residuals vs Fitted Values) 28

Figure 17 Residual Analysis for Logistic Regression (Q-q plot) 29

Figure 18 Residual Analysis for Logistic Regression (Leverage plot) 29

**Introduction**

Blood Donation and Data Science

The first volunteer blood donation service was opened in 1921 since then the outreach for donors has been constant. The importance of blood donation is clear given that, there is no known substitute for human blood and its products and blood transfusions are a necessity for many patients ranging from someone who is fighting cancer, a sickle cell disease, a premature infant to someone that has suffered an accident. Blood donation stumbles upon several challenges. Its main challenge is that only 37% of this country’s population is eligible for donation and only 10% of this group are donating annually. The second biggest challenge is their shelf life, red blood cells need to be utilized within 35 to 42 days after donation to ensure the patient's safety. One of the ways to minimize the degree of impact these challenges present is to find a way to accurately predict blood donations. Data Science can utilize predictive analytics to process a given amount of information and analyze it to predict future blood donations. By predicting future blood donations, the healthcare system can better prepare and manage their resources. Blood donation organizations for example can organize staff schedules according to predicted donations as well as to present a stronger campaign for the lower donation months. On the other side of the spectrum the hospitals can better manage their resources by knowing when are they going to become available. Blood donation prediction models can benefit blood donation organizations as well as the patients in need.

**Literature**

In order to better understand the challenges ahead we have chosen three research articles: “Big data analytics in healthcare: promise and potential” by Willianallur Raghuoathi and Viju Raghuoathi, “Prediction of hemoglobin levels in whole blood donors: how to model donation history” by A. Mireille, Yvonne Vergouwe, Femke Atsma, Karel G. M. Moons and Wim L.A.M. de Kort and finally , “Predicting blood donation intentions and behavior among Australian blood donors: testing an extended theory of planned behavior model” by Barbara M. Masser, Katherine M. White, Melissa K. Hyde, Deborah J. Terry, and Natalie G. Robinson. These articles present arguments of the benefits of data science in the healthcare field as well as methodologies that have been utilized in different healthcare scenarios.

Big data analytics in healthcare: promise and potential

The first article “Big data analytics in healthcare: promise and potential” presents an overview of big data and its increasing role in healthcare, its importance, and methodology. This research argues that healthcare is big data. Healthcare contains voluminous amounts of data with substantial variety going from doctors notes (in physical form) to a patient’s vital signals (machine generated sensor data). The amount and variety of data inside the healthcare system creates a complex but resourceful database that with the right tools it can aid all members of this field. Utilizing Data Science in healthcare has the potential to benefit the patients, doctors, researchers and insurance companies by, providing the users more information regarding specific situation and enabling them to address them in a more successful manner.

The methodology described in “Big data analytics in healthcare...” is presented in four main steps. The first step consists of the development of a “concept statement” (Wullianallur Raghupathi, 2014) which refers to the process of identifying the need, significance, and cost of the project. Step two provides background information regarding the problem as well as the way the team is planning on addressing the problem. The third step breaks down the concept statement into statistical approaches, and starts addressing the case. This step identifies the independent and dependent variables as well as, the collection, and transformation of the data to continue with the analytics face. Once the data analytics techniques have been applied the project enters its fourth face. The final step consists of the testing of the created models and their explanations. In this step data scientists ensure the accuracy of their predictive models if the models and their explanations are validated this phase concludes the project.

Data scientist have the tools to transform healthcare. Healthcare data will only continue to grow and will always contain many forms of data if, this data is analyse it can not only provide healthcare professionals with the knowledge necessary to save a life but also reduce the cost of this process. Data science in the healthcare system is starting to lay down roots for a greater stage of development which will provide platforms and tools necessary for providing better healthcare.

Prediction of hemoglobin levels in whole blood donors: how to model donation history

The second article “Prediction of hemoglobin levels in whole blood donors: how to model donation history” tackles the danger of the probability of recurrent blood donors presenting low hemoglobin (Hb) levels after a previous donation. Hemoglobin level directly relates to range if iron in human blood. Blood donation organizations ensure that donors present the right hemoglobin range required for blood donation to prevent the donor from suffering anemia or give iron deficient blood to a patient. This study focuses on generating a prediction model of “whole blood donors” meaning not plasma donors, with possible low levels of Hb. The goal of this study is to predict blood donations of loyal donors by measuring their Hb level before and after donation.

This research utilized linear regression model containing the predictor mean of all previous hemoglobin levels of the loyal donors. They concluded that this model was more accurate for both sexes given that it contains more data than a model with only the pre donation hemoglobin data set. The researchers also created several other models but realized that they usually fitted female or male donors, which they regarded as less effective compared to the linear regression models which applies to both sexes.

By predicting the hemoglobin levels for loyal blood donors, blood donation organizations can accurately predict how many candidates for donation are going to have the correct hemoglobin levels. This information aids blood banks and blood donation organizations to predict the amount of supplies, from loyal donors, throughout the year.

Predicting blood donation intentions and behavior among Australian blood donors: testing an extended theory of planned behavior model

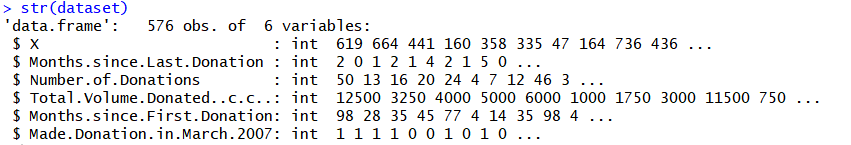
The third article, “Predicting blood donation intentions and behavior among Australian blood donors: testing an extended theory of planned behavior model” talks about the diverse intentions and behaviors among blood donors. The intentions and behaviours discussed by the authors of the paper include: “Attitude, Subjective norm, Self-efficacy, Intention, Moral norm, Self identity, Anticipated regret, Donation anxiety, Behaviour and Demographic details” (Masser BM, 2008). The data on these predictor variables leading to blood donor forecasting was collected using a questionnaire which was distributed to 263 students from college students who participated in blood donation in Queensland, Australia.

The authors initialize this project by analysing the correlational relationships between the potential predictors (attitude, subjective norm, self-efficacy, intention, moral norm, self identity, anticipated regret, donation anxiety, behaviour and demographic details) with both intention and behavior. They identified the strongest correlations followed by the application if the multiple regression model. The model generated by this method was estimated using a robust weighted least-squares estimator. In order to test the validity of the model a number of goodness-of-fit indices were calculated, along with Chi-square, Root Mean Square Error of approximation (RMSEA), Weighted Root Mean Square residual(WRMR) and a Comparative Fit Index(CFI)l. The final outcome indicated that chi-square statistics was non significant and the CFI was above 0.95. In addition, the authors had a RMSEA below 0.08 and WRMR below 0.90. The authors concluded that the predictors were significantly correlated with each other. They also identified various predictors that were negatively correlate with other ones.

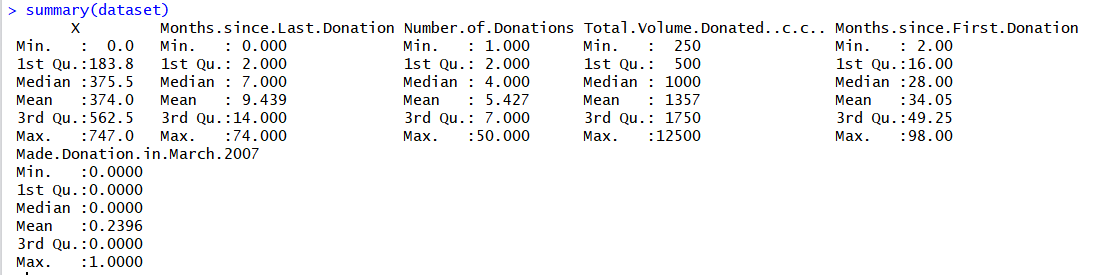
**Data Mining/Cleaning**

For this project, there was no need to do any kind of data cleansing technique. All the data rows contained complete set of values, indicating the inexistence of null values. Given this conditions the data was utilized as provided by Drivendata.org. There was, however, the addition of three new variables into the original data to aid with the prediction of the blood donation pattern.

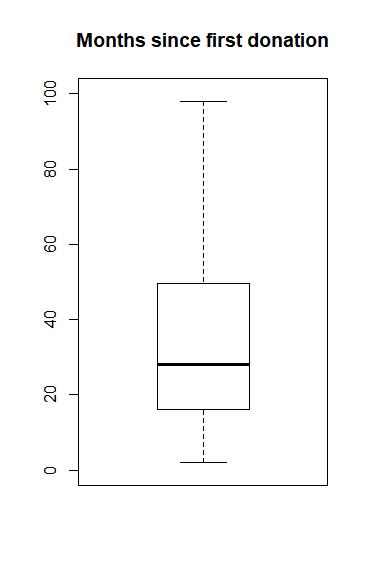
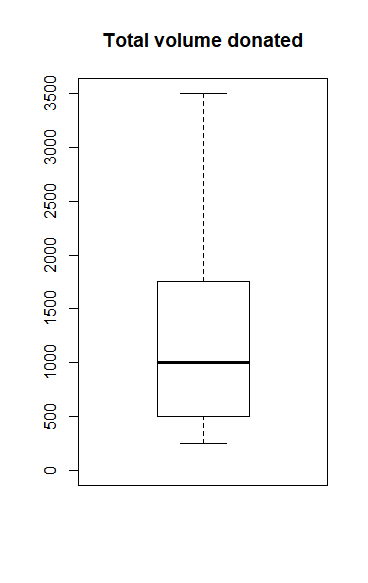
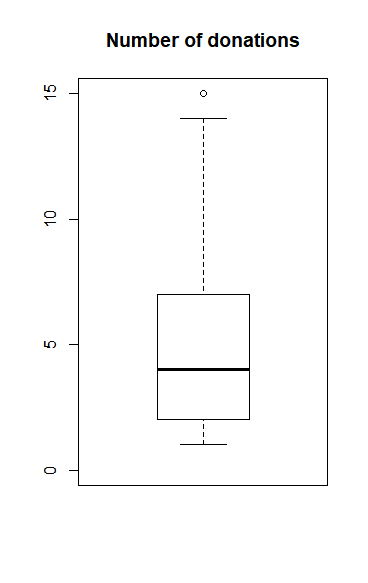
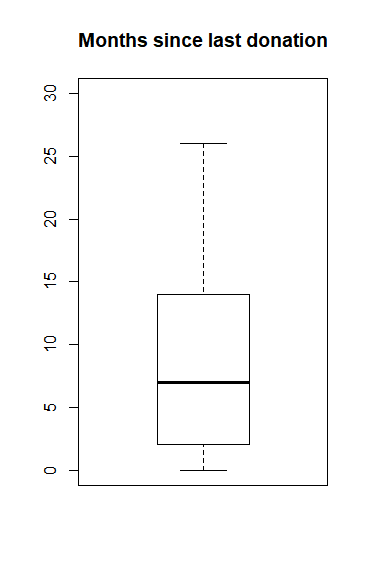
The initial look at the data set revealed the following structure:



The initial summary statistics of the data set showed the following:



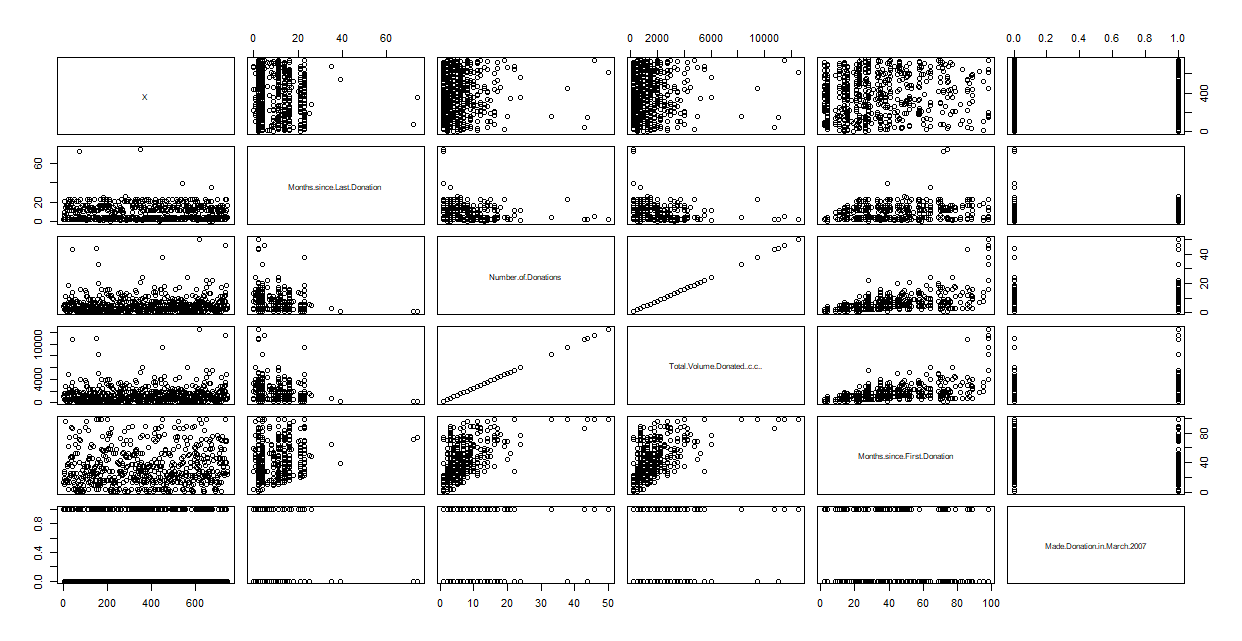
The five number summary of the above data can be shown using a boxplot:

Figure 1 Boxplots for possible predictors

The initial investigation into the dataset revealed multicollinearity between “Number of donations” and “Total Volume Donated”. Given this condition the variable “Total Volume Donated” was removed given that it was unable to act like an independent variable for the analysis.

The prediction analysis was continued by the addition of three more variables. The added variables are: average donations per period, the ratio between “Month since Last donation” to “Month since first donation” (ML-MF Ratio) and a variable z. The variable z represents the ratio of the average donations per period and the difference between the “Month since last donation” and “Month since first donation”.

**Data Visualization**

Figure 2 Graph showing the relationship within the original data

From Figure 2, we can see that “Number of Donations” have perfect correlation with the “Total Volume Donated” variable, which was already expected to have a positive linear relationship. As the “Number of Donations” increases, the “Total Volume Donated” also increases. If the “Total Volume Donated” variable is divided by the “Number of Donations” for each person, we get a constant value of 250 cubic centimeters (cc) which indicates that each person is required to donate 250cc of volume each time. Since these 2 values are dependent on each other, they were not used as independent variables for the created models. These circumstances encourage the decision of dropping the “Total Volume Donated” variable and keping “Number of Donations” as an independent variable for the models.

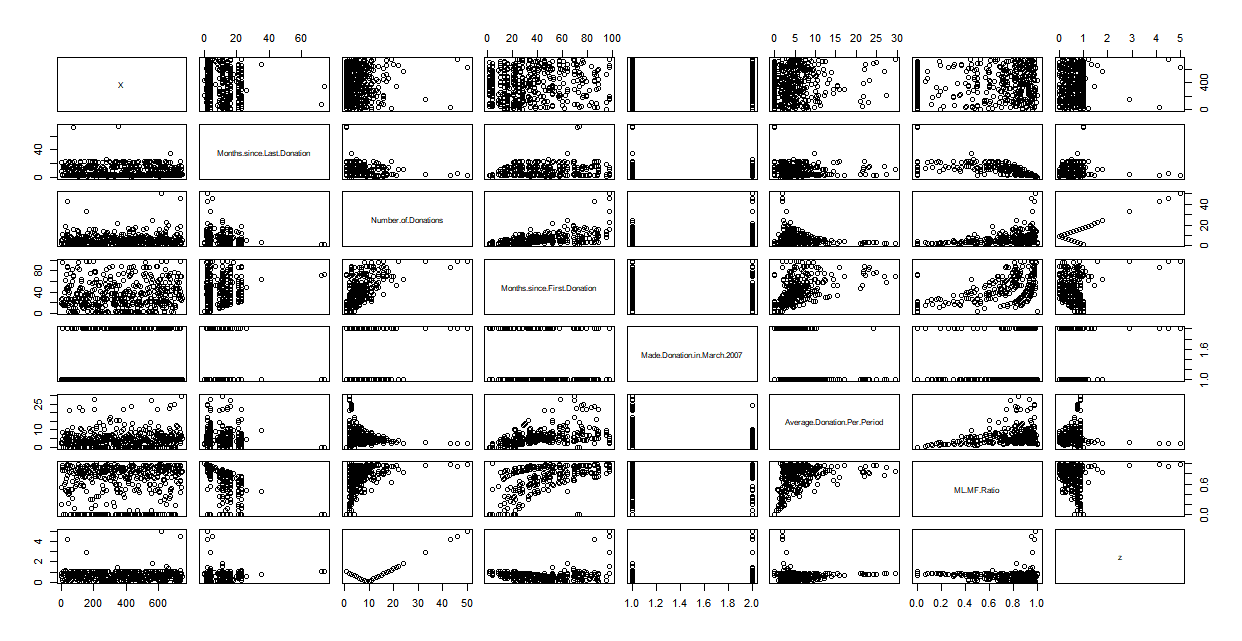
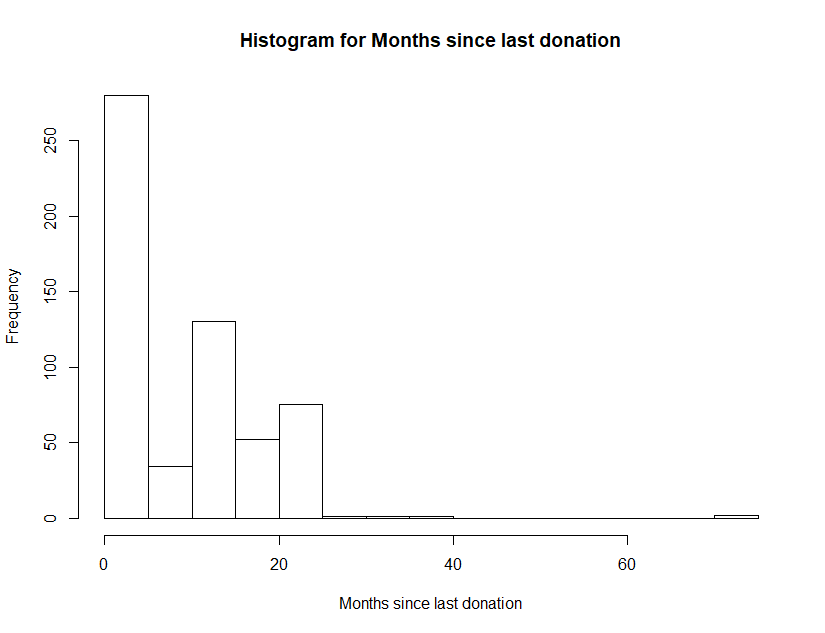
Figure 3 Graph showing the relationship with the new added variables

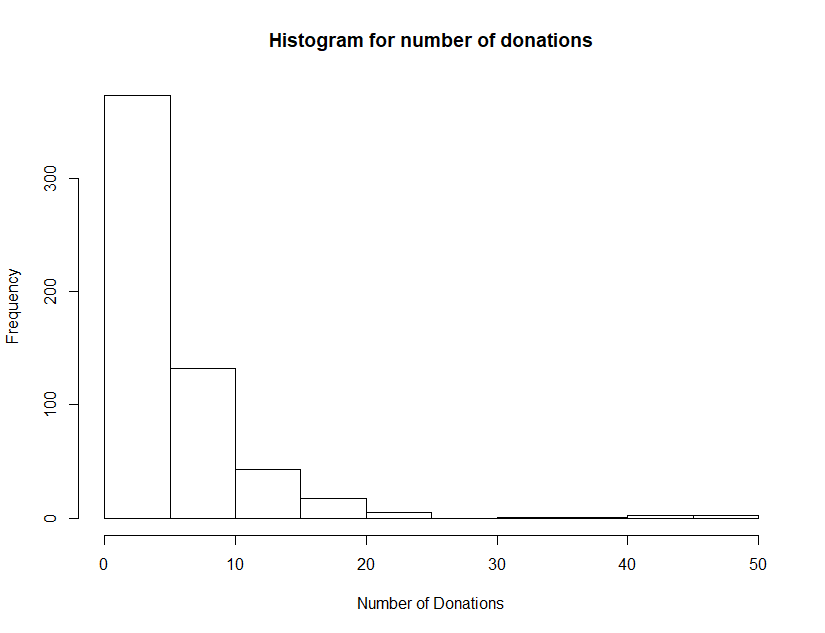
Figure 3, shows the positive correlation between the “Number of Donations” and “Months since first donation”. Similarly, we can see a significant positive correlation between “Month since first donation” and the “average donations per period”, continuing with ‘average donation per period” and ML-MF Ratio, and finalizing with the positive correlation between “Month Since First Donation” and ML-MF Ratio.

Univariate Visualization of the Variables

In this section, all the variables from the data have been visualized with the purpose of identifying the existence of normal distribution. The first step was to plot the “Number of Donations” and “Months Since Last Donation” variables into a histogram. The plot indicated large correlation effects; in which the dependent variable “Made Donation in March 2007” has a small negative correlation with both “Months Since Last Donation” and “Months since First Donation” according to the plots created.

Figure 4 Histogram for Months since last donation

From the above histogram, we can see that the most frequent donors return within a few months after donation, making them the greatest contributors to the “Total Volume Donated”. In other words, this could indicate that people who did not donate for 7 months or more, past their previous donation, are more likely to not be recurrent donors. Which indicates the opposite for the segment of the population meaning that people who frequently donate blood are more likely to return than the ones who don’t donate frequently.

Figure 5 Histogram for number of donations

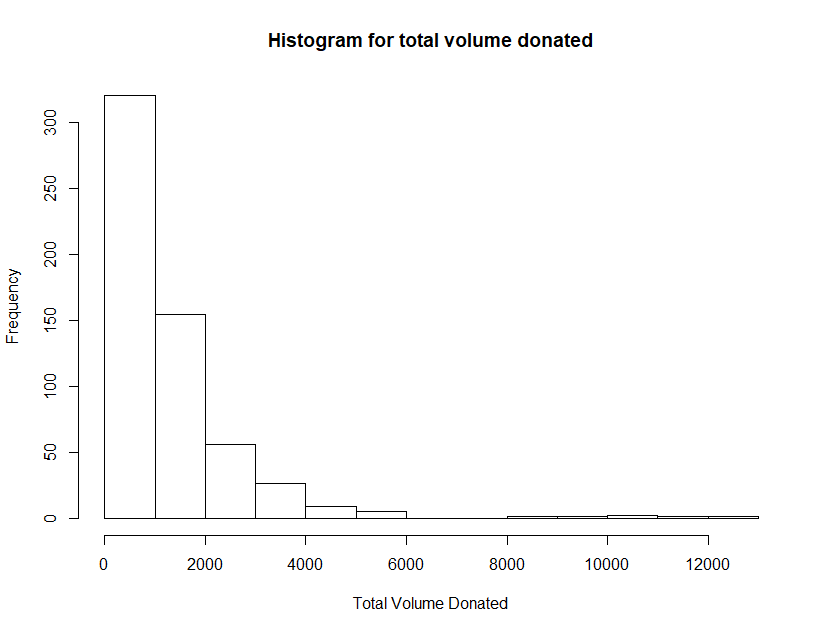
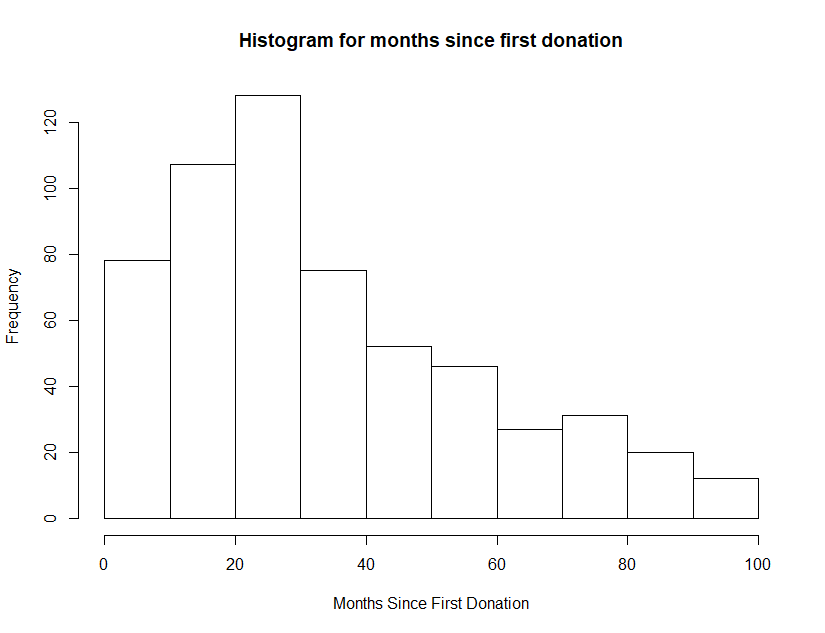
The above histogram shows that the majority of population donate less than 10 times. Indicating that “loyal” donors represent a small portion of the population in the data set.

Figure 6 Histogram for total volume donated

The above histogram shows that people that have donated less than 2000 cc of blood are more in numbers compared to the one donating the blood more than 2000 cc. Which is in accordance to the previous histogram, Figure 5.

Figure 7 Histogram for Months since first donation

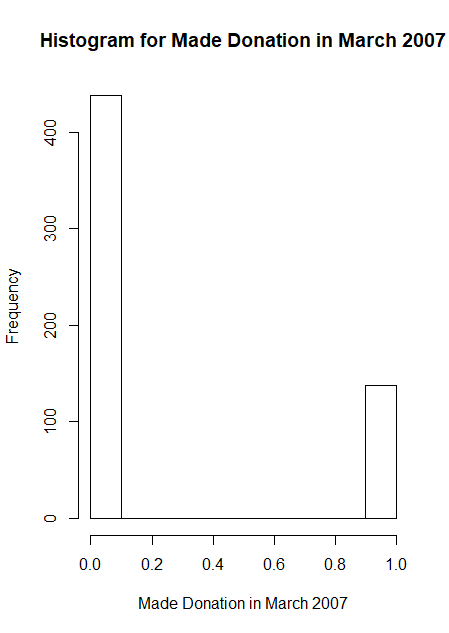
The above histogram shows that people between 20 to 30 months since their first blood donation seem to donate more.

Figure 8 Histogram for Made donation in March 2007

The above histogram shows that the proportion of people who didn’t donated blood in March 2007 is more than twice the number of people that did.

Looking at the distributions of the dependent and independent variables, led to the conclusion of introducing three more variables that can positively contribute to the predictive model.

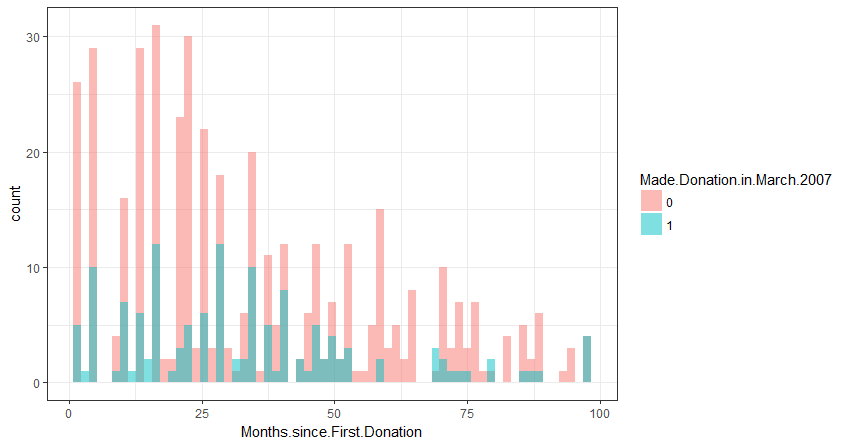
The first variable added is the “average donation per period”. For this variable we take the difference between first donation to last donation for each person and divide that by the total number of times the individual donated blood. This variable gives the unit difference of time for each person’s donation.

The next variable introduces is ML-MF Ratio. This variable gives us the ratio between the “Month since the last donation” to the “Month since the first donation”.

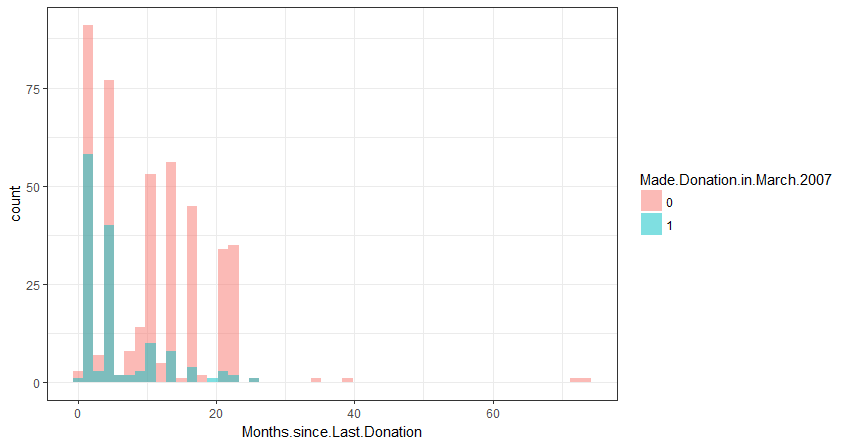
Finally the last variable is z. To calculate the z value represents the subtraction of the sum of the mean of difference between the first donation to last donation time from frequency of donation for each person who donated blood divided by the standard deviation of the difference between the first donation to last donation time per person.

Multivariate Visualization of the Variables

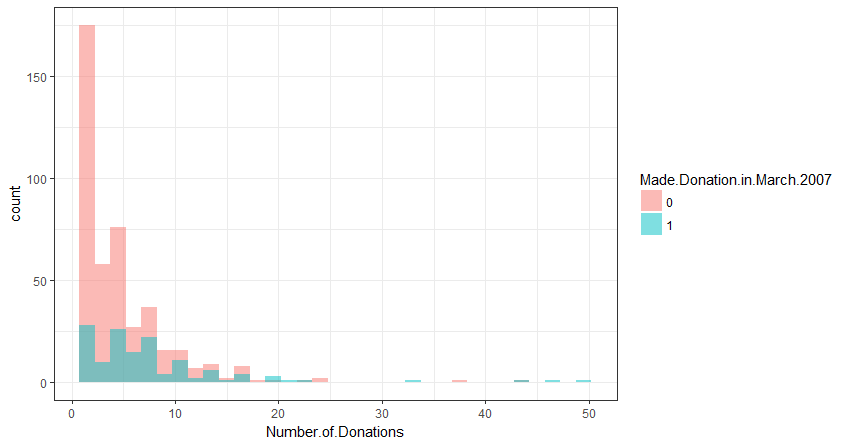
The next section contains multivariate visualization between the independent variables and all the dependent variables.

Figure 9 Multivariate histogram for Months since first donation

In this histogram, the red plots are indicating the number of months since the first donation for each person who did not return to donate blood in March 2007. The blue plots are indicating the number of months since the first donation for each person who returned to donate blood in March 2007.

Figure 10 Multivariate histogram for months since last donation

In Figure 10, the red plots are indicating the number of months since the last donation for each person who did not return to donate blood in March 2007. The blue plots represent the number of months since the last donation for each person who returned to donate blood in March 2007.

Figure 11 Multivariate histogram for number of donations

In the above histogram, the red plots are indicating the number of total donations each person made of which didn’t return to donate blood in March 2007. The blue plots are indicating the number of total donations for each person who returned to donate blood in March 2007.

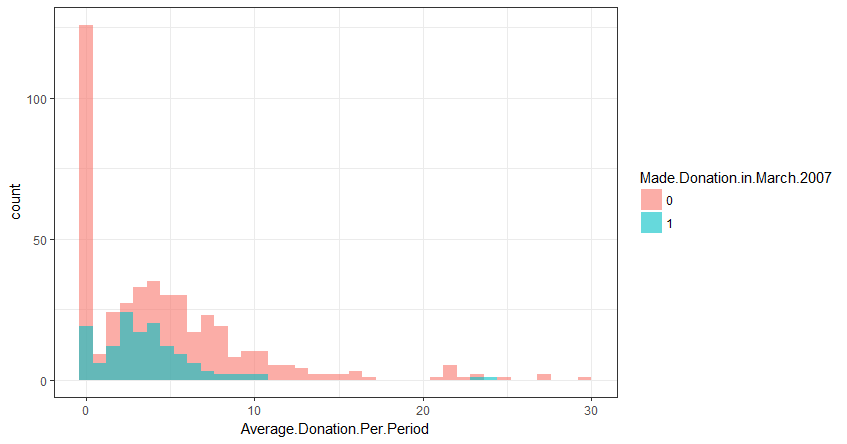
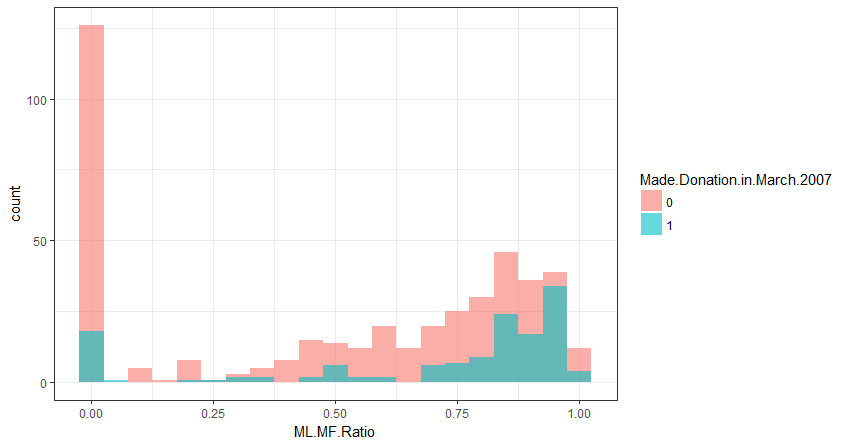
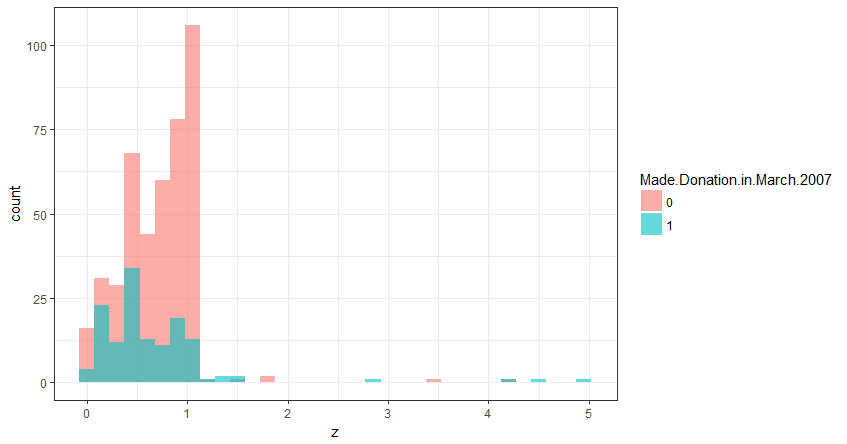


Figure 12 Multivariate histogram for Average donation per period

In this histogram, the red plots are indicating the average donation each person made per period who did not return to donate blood in March 2007. The blue plots are indicating the average donation each person made per period who returned to donate blood in March 2007.

Figure 13 Multivariate histogram for Ml-MF ratio

In this histogram, Figure 12, the red plots are indicating the month since last to first donation for each person who did not return to donate blood in March 2007. The blue plots are indicating the month since last to first donation for each person who returned to donate blood in March 2007.

Figure 14 Multivariate histogram for Made Donation in March 2007

In the above histogram, the red plots are indicating the z value for each person who did not return to donate blood in March 2007. The blue plots are represent the z value for each person who returned to donate blood in March 2007.

**Predictive Analysis**

For the blood donation project, drivendata, provides two types of data sets the “train set” and the “test set” accompanied by a visual example of the submission format. The purpose of this project is to utilize the independent variables to create a prediction model indicating whether a previous donor will donate again. Given that it is necessary to predict the dependent variables, that is a categorical variable, a traditional regression model will not be helpful. For this reason the Logistic Regression and Support Vector Machine (svm) with radial kernel will be used to build the model.

Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome which is measured with two possible values for a variable. The variable in question has only two possible outcomes. I this case, there was a blood donation the value is 1 (TRUE, success, yes etc.) or 0 indication no donation (FALSE, failure, no etc.).

The goal of logistic regression is to find the best fitting model to describe the relationship between the variable in question (yes or no donation) and a set of independent (predictor or explanatory) variables. Logistic regression generates the coefficients as well as its standard errors and significance levels to predict a logic transformation of the probability.

Rather than choosing parameters that minimize the sum of squared errors, estimation in logistic regression chooses parameters that maximize the likelihood of observing the sample values.

**R Code:**

# Fitting Logistic Regression to the Training set

LR\_classifier = glm(formula = Made.Donation.in.March.2007 ~ Months.since.Last.Donation

+ Number.of.Donations

+ Months.since.First.Donation

+ Average.Donation.Per.Period

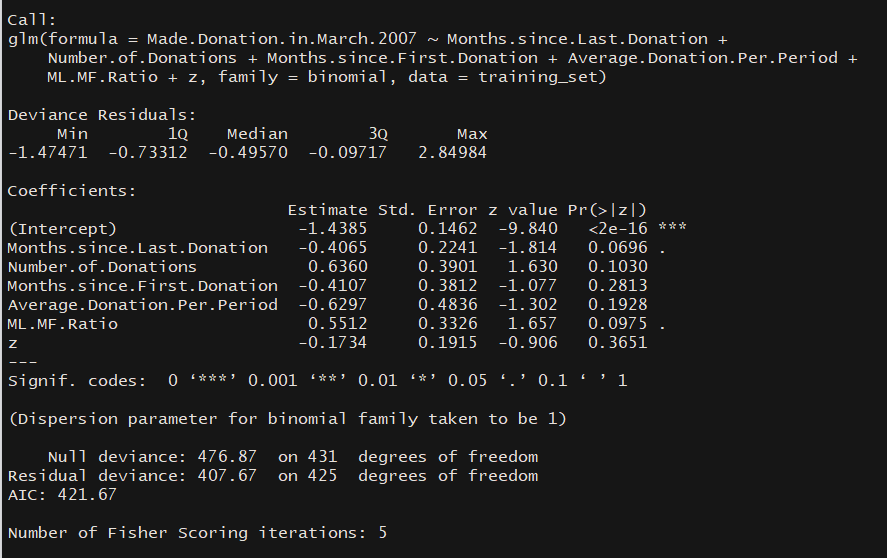
+ ML.MF.Ratio

+ z ,

family = binomial,

data = training\_set)

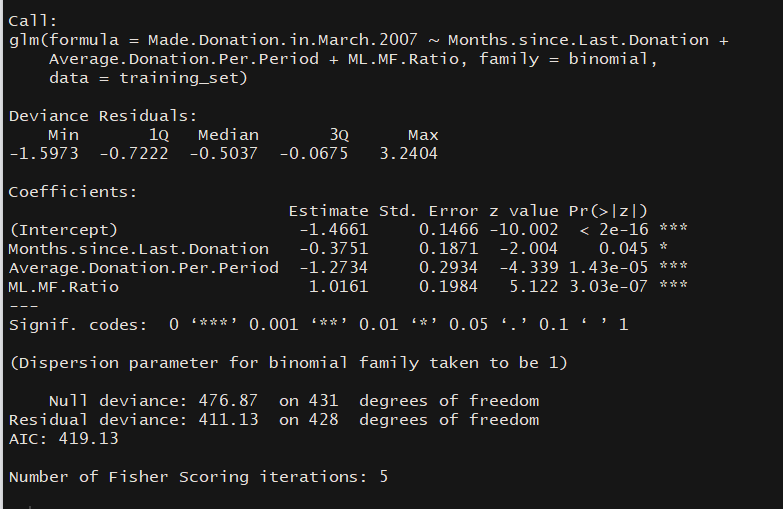
**Output**:



From this output we can see the deviance residuals are not proportionately distributed. Also four of the coefficients are showing negative values which indicates they have negative correlations with the categorical values. But the coefficients are not showing statistical significance since the p-values are larger than 0.05. The stars beside the p-values indicate statistical significance of independent variables in a model which, indicates that by removing some of the independent variables then the model might perform better. Also we should take notice of the null deviance, residual deviance, and AIC values.

[Deviance](http://www.theanalysisfactor.com/concepts-you-need-to-understand-to-run-a-mixed-or-multilevel-model/) is a measure of goodness of fit of a generalized linear model; higher values indicate a worse fit. The null deviance shows how well the response variable is predicted by a model that includes only the intercept. In this case, the null deviance we get from the Logistic Regression model is 476.87 with 431 degrees of freedom. Including the independent variables, the value decreases the deviance to 407.67 points on 425 degrees of freedom, which is a significant reduction in deviance.

By removing the variables with the highest p-values from the previous model and running Logistic Regression model again on the dataset we obtain the following outcome.

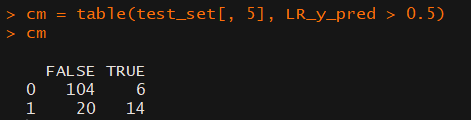


All the p-values of the independent variables have values less than 0.05 and they are statistically significant meaning that, all of these variables contribute significantly to the model. On the other hand the residual variance increased from 407.67 to 411.13 which means the previous model was better in terms of residuals. However, the AIC value decreased from 421.67 to 719.13. Which indicates that this model is better compared to other model.

These two fold answers are very confusing, we couldn’t decide which model to choose, so we tried another method.

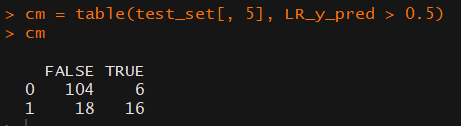
We have took the training data provided by drivendata for the blood donation competition and treated it as a whole dataset. We randomly split the dataset into training and test set keeping the ratio of training data as 0.75. Next logistic regression was performed for both the models described earlier and predicted the values for the test set. The next step was the construction of confusion matrix in order to see which model has a better accuracy.

For the first logistic regression model, all the independent variables were present. Which led to a low residual deviance and high AIC value. Lets see how this model performs with the confusion matrix:

****

Thus the model can predict 118 (= 104 + 14) observations correctly while 26 (= 20 + 6) observations incorrectly. This model gives us an accuracy of approximately 82%.

The second model has less variables included in the model but those variables are statistically significant. This model obtained a lower AIC values but a higher residual deviance. This model performance with the confusion matrix is:



Thus the model can predict 120 (= 104 + 16) observations correctly while 24 (= 18 + 6) observations incorrectly. This model gives us an accuracy of approximately 83.3%.

If we consider the metric used for the competition, the second model should perform better with the log loss function because it has better accuracy in prediction. In conclusion we can state that statistically significant variables give us better accuracy in prediction even though the residual deviance is not lower.

Support Vector Machine

The following model utilized is the Support Vector Machine (SVM) with a ‘radial’ kernel. For this project we needed to implement another model which was not discussed in class. Through some research we encounter 2 models that perform better with categorical data, the Random Forest Classification and SVM. After studying and implementing both these models in the blood donation dataset, we observed that the SVM classifier performed better with our design. This is why, we have decided to discuss more about the SVM classifier.

A Support Vector Machine (SVM) is a discriminative classifier defined by a separating hyperplane. Given labeled training data, the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side [6].

**R Code:**

library(e1071)

svm\_classifier = svm(formula = Made.Donation.in.March.2007 ~ Months.since.Last.Donation + Number.of.Donations

+ Months.since.First.Donation + Average.Donation.Per.Period

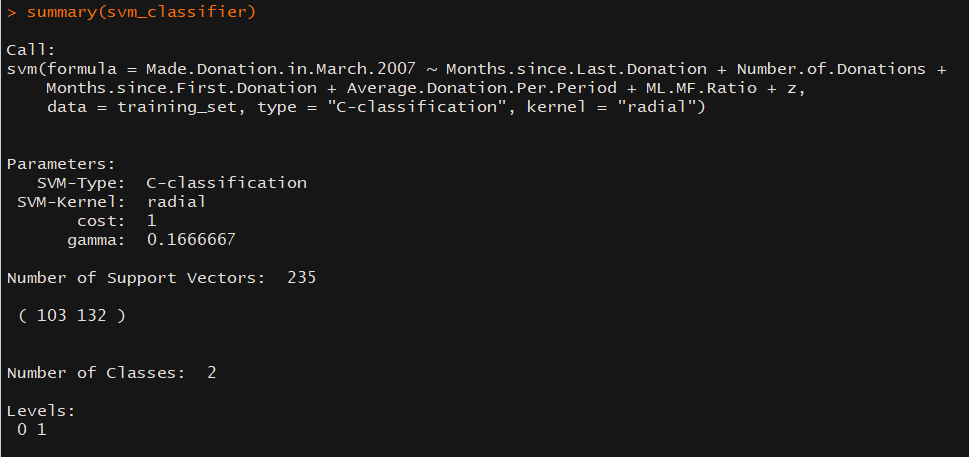
+ ML.MF.Ratio + z ,

data = training\_set,

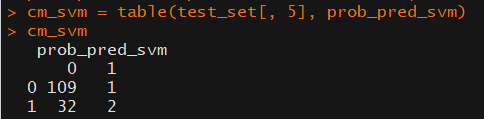
type = 'C-classification',

kernel = 'radial')

**Output:**

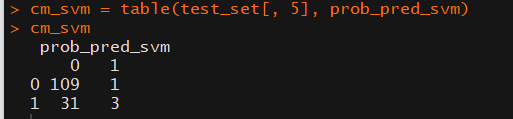


The training data provided by drivendata, for the blood donation competition, was treated it as a whole dataset; opposed to randomly splitting the dataset into training where the training data has a 0.75 ratio. Next we perform 2-class (2 classes are for categorical values 0 and 1) SVM classification to predict the values for the test set. After this we utilized the confusion matrix to compare the accuracy of this model against the previous ones.



In this case, the model can predict 111 (= 109 + 2) observations correctly while 33 (= 32 + 1) observations incorrectly. This model gives us an accuracy of approximately 77% which is much lower than both of the previous Logistic Regression models.

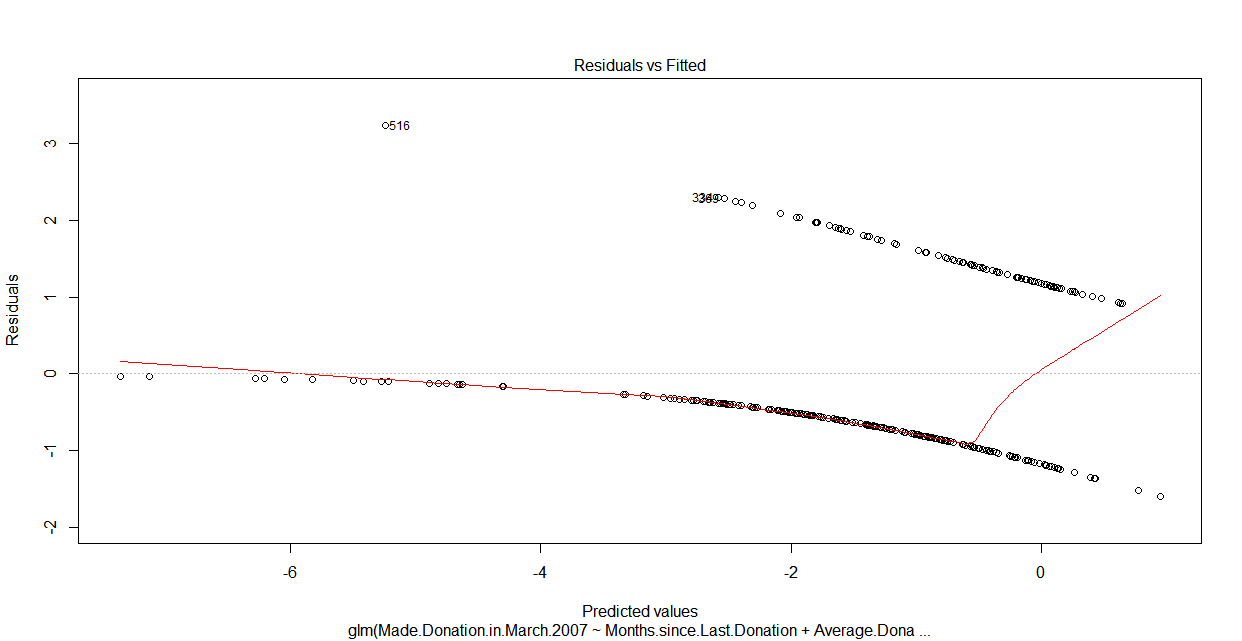
The next step is to perform trial and error method with the SVM model by trying to eliminate the independent variables that are not statistically significant, to the model. Our goal here was to find the combination of independent variables that give us better accuracy with the confusion matrix. By analyzing different models, we have found that the model without the z value gives the best accuracy. The confusion matrix by the model is shown below:



In this case, the model can predict 112 (= 109 + 3) observations correctly while 32 (= 31 + 1) observations incorrectly. This model gives us an accuracy of approximately 77.7% which is much lower than both, previous Logistic Regression models.

**Discussions and Recommendations**

Analysis for Logistic regression model

Figure 15 Residual Analysis for Logistic Regression (Residuals vs Fitted Values)

We can see some heteroscedasticity in our residuals, i.e. the plot of the residuals against the fitted values show some kind of pattern but this is due to our prediction being either 0 or 1.

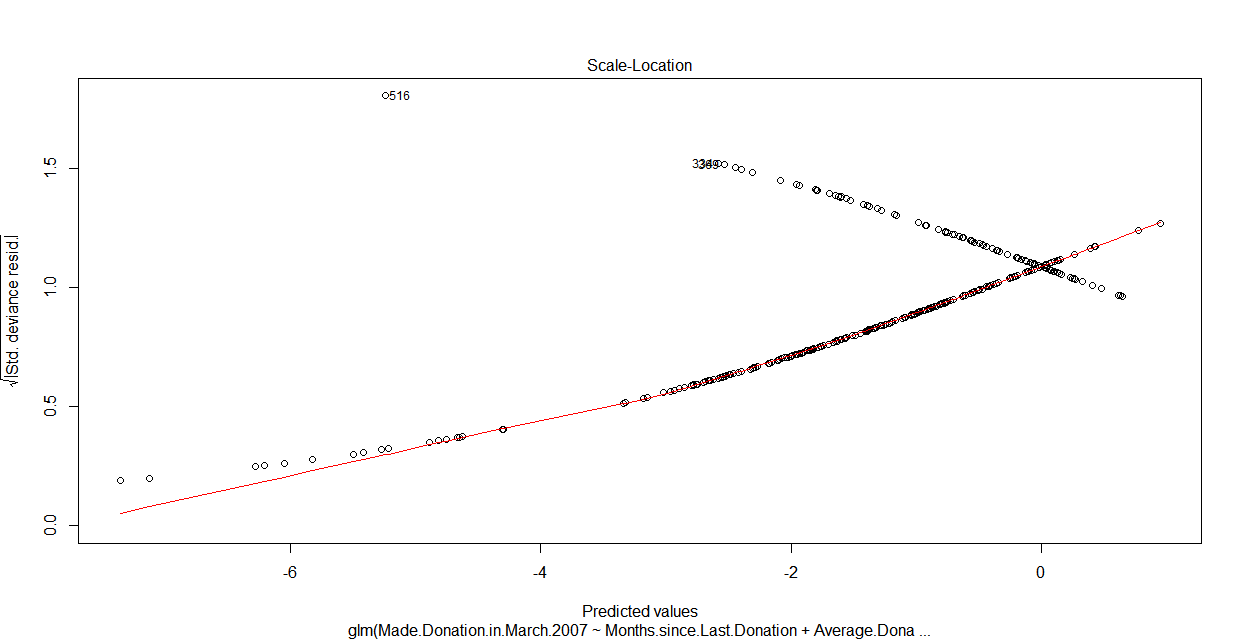
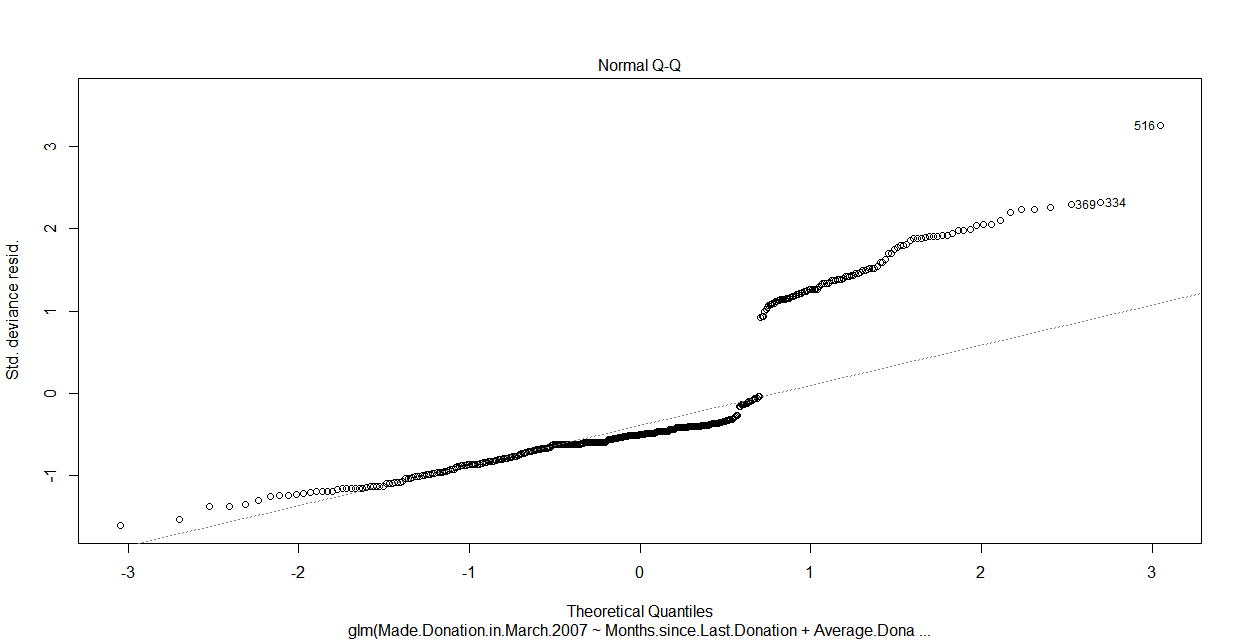
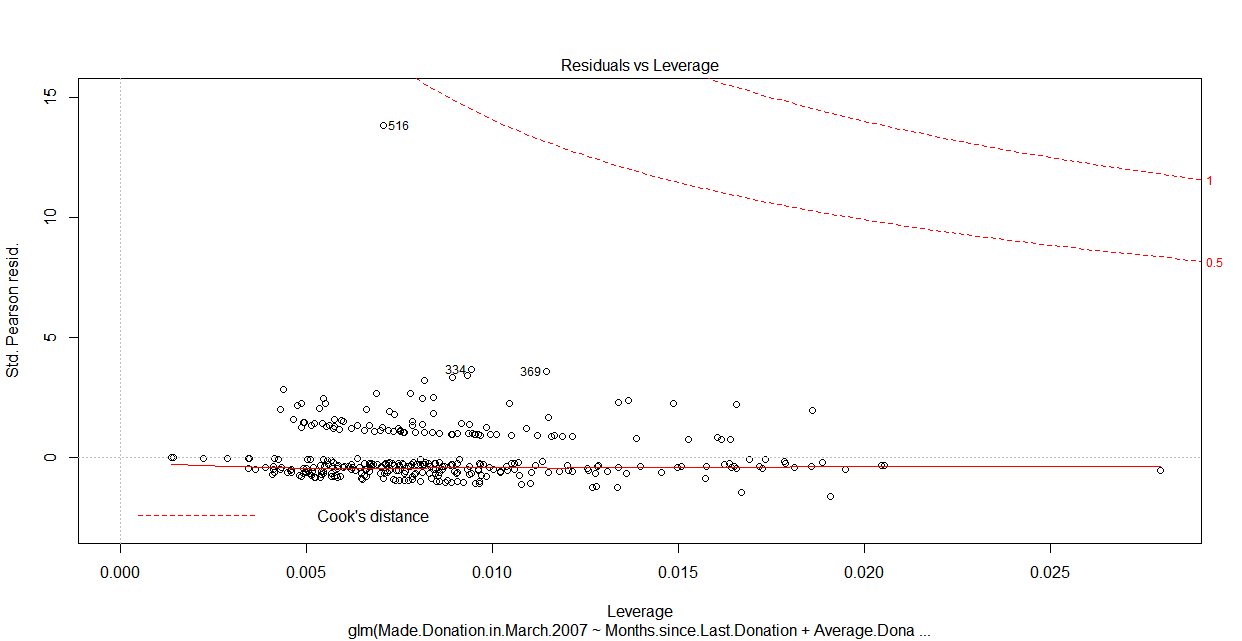
Figure 16 Residual Analysis for Logistic Regression (Standardized Residuals vs Fitted Values)

Figure 16 is quite similar to Figure 15. There are some patterns in this graph, suggesting heteroscedasticity in our residuals.

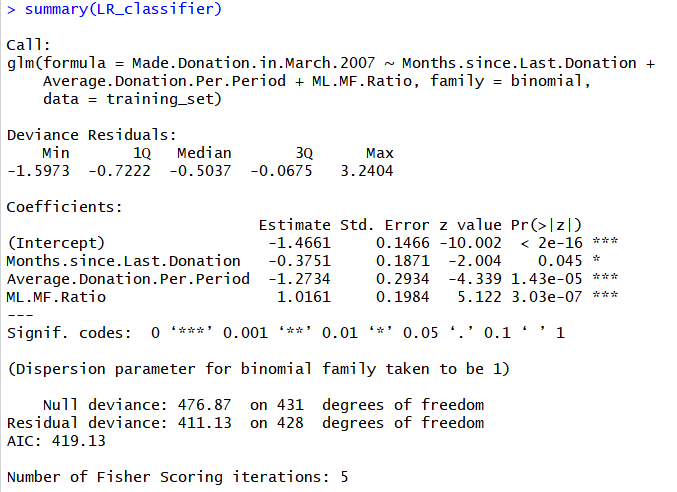
Figure 17 Residual Analysis for Logistic Regression (Q-q plot)

From Figure 17, we can observe that the residuals are not normally distributed.

Figure 18 Residual Analysis for Logistic Regression (Leverage plot)

The plot for leverage shows that all the residuals are however, within the Cook’s distance.

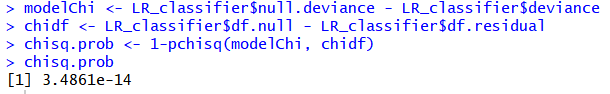
The overall performance of the logistic regression can be evaluated using the Null and Residual deviance.



The Null deviance represents deviance of a model with only the grand mean, so the reduction in deviance represents the deviance explained by the model.

The Chi-square p-value is 3.4861e-14, which shows that our Logistic regression model is good.

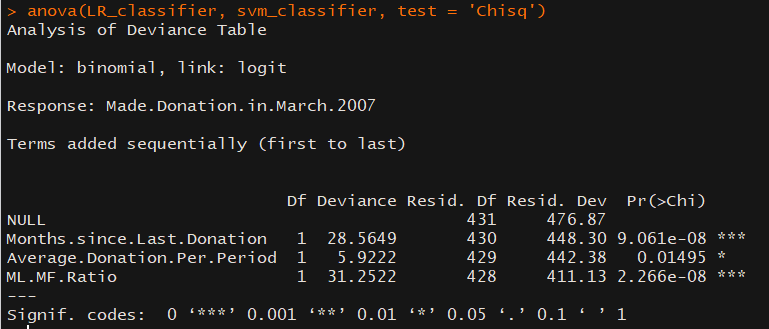
R-code:



Also, from the coefficients of our regression model can in interpreted as log of the change in the odds ratio. The slope of -0.3751 for “Month since last donation”, indicates that each “Month since last donation” decreases the odd of blood donation by 0.3751 folds. Similarly, the slope of average donation (-1.2734), tells us that each increase in the average number of times of a blood donation, the odds of the blood donation decreases by 1.2734 folds and the slope of ML-MF ration (1.0161), tells us that, with per point increase in this ratio, and the odds of blood donation increases by 1.0161 folds.

Comparing the Logistic Regression and SVM Models

Next we perform ANOVA Chi-square test on both of our Logistic Regression and SVM models to check if these models are statistically significantly different or not.



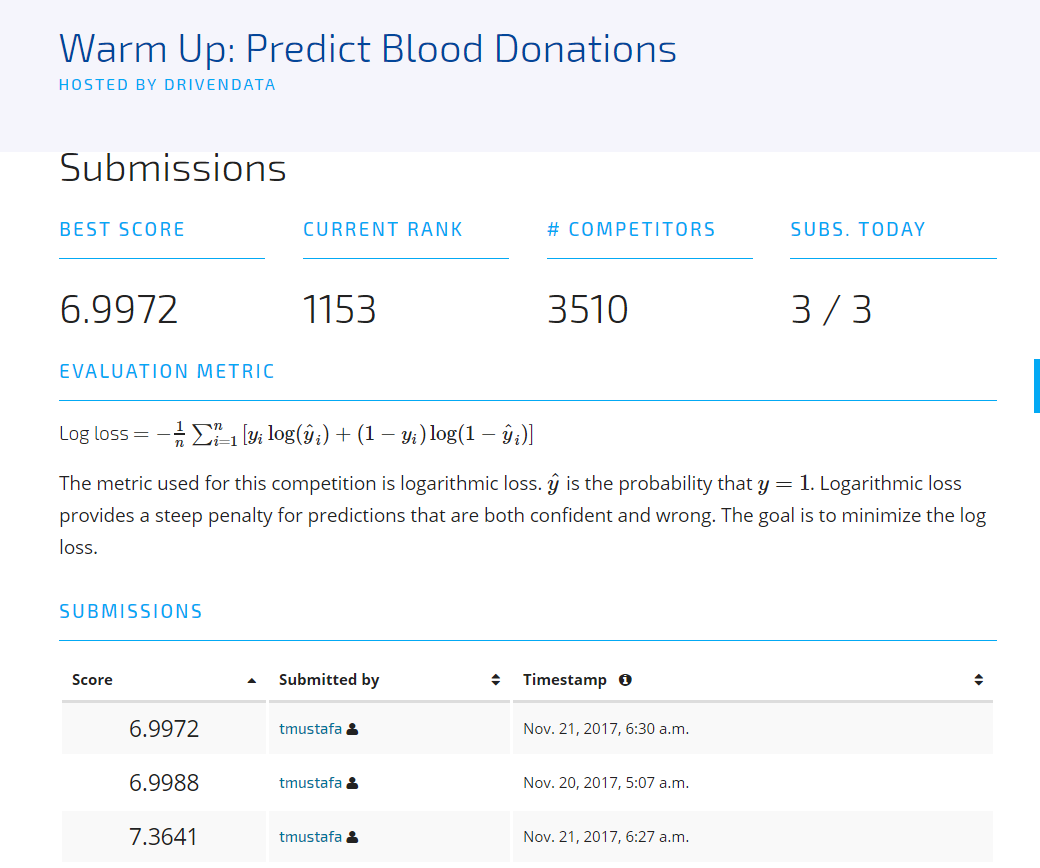
In the anova utilized for the Logistic Regression model, all the p-values of the independent variables are below 0.05; which means that the fitted Logistic Regression model is significantly different from the SVM model at the level of α=0.05.

Recommendation

To model our blood donation pattern, we can improve our current model by including more variables into our model like donor demographics, attitude, subjective norm, self-efficacy, intention, moral norm, self-identity, anticipated regret, donation anxiety and behavior.

In order to improve the number of people donating blood, we can focus more in blood donation campaigns focused towards people who recently donated blood against those people who have donated blood in a long period of time. The reason is that by focusing more on people who have recently donated blood would increases the odds of these individuals to donate their blood once again.

**Driven Data Submission**

****

**References:**

[1] A. Mireille Baart, Y. V. (2013, Sep 30). *Prediction of hemoglobin levels in whole blood donors:*

*how to model donation history*. Retrieved Nov 17, 2017, from Wiley Online Library: The third article, “Predicting blood donation intentions and behavior among Australian blood donors: testing an extended theory of planned behavior model” by Barbara M. Masser, Katherine M. White, Melissa K. Hyde, Deborah J. Terry, and Natalie G. Robinson (September 1, 2008).

[2] GivingBlood. (n.d.). *Why Give Blood*. Retrieved Nov 16, 2017, from Community Blood Center

Real Donors. Real Life Heroes.: <http://givingblood.org/donate-blood/why-give-blood.aspx>

[3] Masser BM, W. K. (2008, Nov 19). *Wiley Online Library*. Retrieved Nov 17, 2017, from

Predicting blood donation intentions and behavior among Australian blood donors: testing an extended theory of planned bahavior model.: <http://onlinelibrary.wiley.com/doi/10.1111/j.1537-2995.2008.01981.x/abstract>

[4] Program, A. S. (n.d.). *Together we can make a difference*. Retrieved Nov 17, 2017, from ASBP

Armed Services Blood Program: <http://www.militaryblood.dod.mil/Donors/about_blood.aspx>

[5] Wullianallur Raghupathi, V. R. (2014, Feb 7). *Big data analytics in healthcare: promise and*

*potential*. Retrieved Nov 17, 2017, from PMC US National Library of Medecine National Institute of Health: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4341817/>

[6] Chapter 2 : SVM (Support Vector Machine) — Theory,

https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72